

## PATENT ABSTRACTS OF JAPAN

(11) Publication number: **10105572 A**(43) Date of publication of application: **24.04.98**

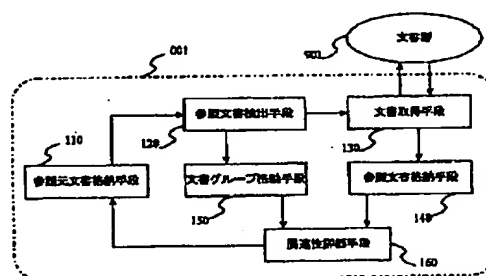
(51) Int. Cl.

**G06F 17/30****G06F 17/21**(21) Application number: **08262047**(71) Applicant: **NEC CORP**(22) Date of filing: **02.10.96**(72) Inventor: **YAMAGUCHI TOMOHARU****(54) DEVICE AND METHOD FOR GROUPING DOCUMENTS****(57) Abstract:**

**PROBLEM TO BE SOLVED:** To determine a collection object range limiting the range to follow the reference relation so as to collect documents having a semantically deep relationship when collecting documents based on the reference relation between documents.

**SOLUTION:** A reference document detecting means 120 extracts a document stored in a reference source document storage means 110, detects the reference relation with the other document out of this document, and stores the completely detected source document in a document group storage means 150. A document possessing means 130 possesses the document, which is detected by the reference document detecting means 120, having the reference relation from a document group and stores this document in a reference document storage means, and a relationship evaluating means 140 evaluates the relationship between the reference document stored in the reference document storage means 140 and the document group stored in the document group storage means 150. When there is any relationship, the reference document is added to the reference source document storage means 140 as a new reference source document, and operation is repeated from processing at the reference document detecting means 120. When there is no document stored in the reference source document storage means 120, it is determined the documents stored in the document storage means 130 belong to one group.

COPYRIGHT: (C)1998,JPO



(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開平10-105572

(43)公開日 平成10年(1998)4月24日

(51)Int.Cl.<sup>6</sup>

識別記号

FI

G 0 6 F 17/30  
17/21

G 0 6 F 15/401  
15/20  
15/40

3 1 0 D  
5 7 0 N  
3 7 0 A

審査請求 有 請求項の数12 OL (全 9 頁)

(21)出願番号 特願平8-262047

(22)出願日 平成8年(1996)10月2日

(71)出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72)発明者 山口 智治

東京都港区芝五丁目7番1号 日本電気株式会社内

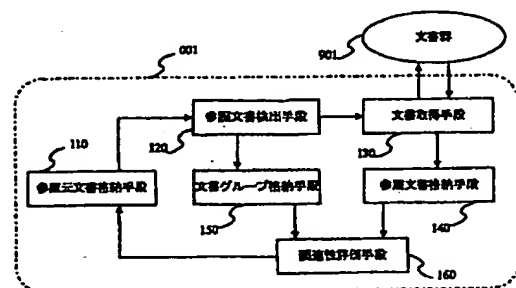
(74)代理人 弁理士 京本 直樹 (外2名)

(54)【発明の名称】 文書グループ化装置および文書グループ化方法

(57)【要約】

【課題】 文書の参照関係に基づいた文書の収集において、意味的に関連性の深い文書を収集できるように参照関係を辿る範囲を限定した収集対象範囲を決定する。

【解決手段】 参照文書検出手段は、参照元文書格納手段に格納された文書を取り出して該文書中から他の文書への参照関係を検出し、検出し終えた参照元文書を文書グループ格納手段に格納する。文書取得手段は、参照文書検出手段により検出された参照関係のある文書を文書群から取得して参照文書格納手段に格納し、関連性評価手段は、参照文書格納手段に格納された参照文書と文書グループ格納手段に格納された文書群の関連性を評価し、関連がある場合には参照文書を参照元文書格納手段に新たな参照元文書として追加し、参照文書検出手段の処理から繰り返す。参照元文書格納手段に格納されている文書が無くなれば、文書格納手段に格納されている文書を一つのグループに属するものと決定する。



## 【特許請求の範囲】

【請求項1】文書間の参照関係情報を持ち、分散して存在する文書群について、分散された文書を収集する際に、任意の特定の文書を起点として、該文書から参照関係を辿って到達可能な文書群のうち、該文書に関連の深い文書のみを効率よく収集するために、参照関係に基づいて参照される文書と、収集済みの文書群との関連性を判定することにより、参照される文書を収集対象とするか否かを決定することで、参照関係を辿る範囲を限定することを特徴とする文書グループ化装置。

【請求項2】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化装置において、ある文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報を取り出して、前記参照関係情報に基づいた参照文書を収集し、

前記参照元文書と前記参照文書の関連性を判定し、前記参照文書が前記参照元文書と関連性がある場合には、前記参照文書を参照元文書に追加し、さらに追加された参照元文書から参照可能な文書の関連性を判定することによって、ある文書に関連性のある文書を収集して、グループ化することを特徴とする文書グループ化装置。

【請求項3】分散して存在する複数のハイパーテキスト文書の中から、あるハイパーテキスト文書に関連性のある他のハイパーテキスト文書を収集して、関連性があるハイパーテキスト文書同士をグループ化する文書グループ化装置において、

前記ハイパーテキスト文書中のリンクを順次辿って関連文書を収集する際に、リンク先の文書と収集済みの文書との関連性を判定することにより、リンク先の文書を収集対象とするか否かを決定し、収集対象としない場合には、そのリンク先の文書からのリンクも辿らないことにすることで、あらかじめ特定の収集条件を設定することなくリンクを辿る範囲を限定することを特徴とする文書グループ化装置。

【請求項4】HTML形式で記述されたハイパーテキスト文書が、ネットワークを介して複数の計算機内に存在し、ある特定のハイパーテキスト文書に関連性が有る文書を前記計算機から収集して、関連性のある文書同士をグループ化する文書グループ化装置において、文書収集の起点となるハイパーテキスト文書（参照元文書）から、他のハイパーテキスト文書の参照を示すURLを検出し、ネットワークを介して前記URLに該当するハイパーテキスト文書（参照文書）を収集し、前記参照元文書と収集された前記参照文書の関連性が有るか無いかを判断し、前記参照文書が関連性が有りと判断されたものは、前記参照文書を参照元文書として追加し、さらに追加された参照元文書内のURLを検出して、他の参照文書を収集する動作を繰り返すことにより、関連性

のあるハイパーテキスト文書をグループ化することを特徴とする文書グループ化装置。

【請求項5】請求項4に記載された文書グループ化装置において、

前記参照元文書と収集された前記参照文書の関連性が有るか無いかを判断する時に、前記参照元文書と前記参照文書に含まれるキーワードを抽出し、前記キーワードの一致度により、関連性を判断することを特徴とする文書グループ化装置。

10 【請求項6】文書収集の起点となる文書を格納する参照元文書格納手段と、

前記参照元文書格納手段に格納された文書を順次取り出して該文書中から他の文書への参照関係を記述した箇所を検出する参照文書検出手段と、

前記参照文書検出手段により検出された参照関係により、前記参照関係に対応する文書を文書群から取得する文書取得手段と、

前記文書取得手段により取得された文書を格納しておく参照文書格納手段と、

20 前記参照文書検出手段により文書中の参照関係を検出し終えた参照元文書を参照元文書格納手段から移して格納しておく文書グループ格納手段と、

前記参照文書格納手段に格納された参照文書と文書グループ格納手段に格納された文書群の関連性を評価し、関連がある場合には参照文書を参照元文書格納手段に新たな参照元文書として追加する関連性評価手段とを含んで構成され、

30 文書の参照関係に基づいた文書の収集において、到達可能な全文書を探索することなく意味的に関連性の深い文書を収集できるように、参照関係を辿る範囲を限定した収集対象範囲を決定をする文書グループ化装置。

【請求項7】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化方法において、

参照元文書格納手段に格納している文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報を取り出す第1のステップと、

40 前記参照元文書を文書グループとして文書グループ格納手段に格納する第2のステップと、

前記第1のステップにより取り出された参照関係情報により、前記文書群から参照文書を取得する第3のステップと、

前記第2のステップで格納された文書グループの参照元文書と、前記第3にステップにより取得された参照文書との関連性の有りを判断する第4のステップと、

前記第4のステップにより関連性が有りと判断された参照文書を、参照元文書として前記参照元文書格納手段に追加する第5のステップと、

50 前記参照元文書格納手段に、参照関係情報が取り出され

ていない参照元文書が有るか無いかを判断し、参照元文書が有る場合には前記第1のステップに戻り一連の動作を繰り返し、参照元文書が無い場合には、得られた文書グループによりグループ化を決定する第6のステップと、

を備えることを特徴とする文書グループ化方法。

【請求項8】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化プログラムを記録した記録媒体において、

第1の記憶手段内に格納している文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報を取り出す第1のステップと、前記参照元文書を文書グループとして第2の記憶手段に格納する第2のステップと、

前記第1のステップにより取り出された参照関係情報により、前記文書群から参照文書を取得する第3のステップと、

前記第2のステップで格納された文書グループの参照元文書と、前記第3にステップにより取得された参照文書との関連性の有り無しを判断する第4のステップと、

前記第4のステップにより関連性が有りと判断された参照文書を、参照元文書として前記第1の記憶手段に追加する第5のステップと、

前記第1の記憶手段に、参照関係情報が取り出されていない参照元文書が有るか無いかを判断し、参照元文書が有る場合には前記第1のステップに戻り一連の動作を繰り返し、参照元文書が無い場合には、得られた文書グループによりグループ化を決定する第6のステップと、を少なくとも備えるプログラムを格納したことを特徴とする文書グループ化プログラムを記録した記録媒体。

【請求項9】文書収集の起点となる文書を格納する参照元文書格納手段と、

前記参照元文書格納手段に格納された文書を順次取り出して該文書中から他の文書への参照関係の説明を記述した箇所を検出する参照文書検出手段と、

前記参照文書検出手段により文書中の参照関係の説明を検出し終えた参照元文書を前記参照元文書格納手段から移して格納しておく文書グループ格納手段と、

前記参照文書検出手段により検出された参照関係の説明に該当する文書中の記述と文書グループ格納手段に格納された参照元文書の本文の記述とから、前記参照元文書と参照関係のある文書との関連性を評価する関連性評価手段と、

前記関連性評価手段により関連があると評価された文書のみを文書群から取得する文書取得手段と、

前記文書取得手段により取得された文書を格納しておく参照文書格納手段と、から構成され、

前記関連性評価手段は、前記参照文書格納手段が格納した文書を参照元文書として、前記参照元文書格納手段に格納する処理を更に行うことを特徴とする文書グループ

化装置。

【請求項10】HTML形式で記述されたハイパーテキスト文書が、ネットワークを介して複数の計算機内に存在し、ある特定のハイパーテキスト文書に関連性のある文書を前記計算機から収集して、関連性のある文書同士をグループ化する文書グループ化装置において、文書収集の起点となるハイパーテキスト文書（参照元文書）から、他のハイパーテキスト文書（参照文書）を示すURLを検出し、

前記URLを説明している文字列のキーワードと、前記参照元文書の本文中のキーワードの一致度を算出することによって、前記参照元文書と前記参照文書の関連性が有るか無いかを判断し、関連性が有ると判断されたURLで示される参照文章を前記ネットワークを介して該計算機から得ることによって関連性のある文書を収集し、さらに、得られた参照文章を参照元文書として他の参照文書を収集する動作を繰り返すことによって、前記参照元文書に関連性が有る文書同士をグループ化することを特徴とする文書グループ化装置。

【請求項11】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化方法において、

参照元文書格納手段に格納している文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報の説明を取り出す第1のステップと、

前記参照元文書を文書グループとして文書グループ格納手段に格納する第2のステップと、

前記第1のステップにより取り出された参照関係情報の説明の内容と、前記参照元文書の本文の内容との関連性の有り無しを判断する第3のステップと、

前記第3のステップにより関連性が有りと判断された参照文書を、前記文書群から取得する第4のステップと、

前記第4のステップにより取得された参照文書を、参照元文書として前記参照元文書格納手段に追加する第5のステップと、

前記参照元文書格納手段に、参照関係情報が取り出されていない参照元文書が有るか無いかを判断し、参照元文書が有る場合には前記第1のステップに戻り一連の動作を繰り返し、前記参照元文書が無い場合には、得られた文書グループによりグループ化を決定する第6のステップと、

を備えることを特徴とする文書グループ化方法。

【請求項12】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化プログラムを記録した記録媒体において、

第1の記憶手段内に格納している文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報を取り出す第1のステップと、前記参照元文書を文書グループとして第2の記憶手段に

格納する第2のステップと、  
前記第1のステップにより取り出された参照関係情報の説明の内容と、前記参照元文書の本文の内容との関連性の有り無しを判断する第3のステップと、  
前記第3のステップにより関連性がありと判断された参照文書を、前記文書群から取得する第4のステップと、  
前記第4のステップにより取得された参照文書を、参照元文書として前記第1の記憶手段に追加する第5のステップと、  
前記第1の記憶手段に、参照関係情報が取り出されていない参照元文書があるか無いかを判断し、参照元文書がある場合には前記第1のステップに戻り一連の動作を繰り返し、前記参照元文書が無い場合には、得られた文書グループによりグループ化を決定する第6のステップと、  
を少なくとも備えるプログラムを格納したことを特徴とする文書グループ化プログラムを記録した記録媒体。

#### 【発明の詳細な説明】

##### 【0001】

【発明の属する技術分野】本発明は電子化された文書の収集装置に関し、特に分散された複数の文書を関連性のあるものとないものに区別して収集範囲を決定し、関連性のある文書を収集してグループ化を行う文書グループ化装置および文書グループ化方法、さらに、文書グループ化を行うプログラムを記録した記録媒体に関する。

##### 【0002】

【従来の技術】従来、分散された文書間の参照関係情報をもつ文書群の代表的なものとしてハイパーテキストが知られている。ハイパーテキストは複数の文書をリンクによって関連付けることが可能である。

【0003】この代表的な従来技術の例として、特開平4-321144号公報に記載の「ハイパーテキストのブラウジング処理装置」が知られている。この従来技術によれば、文書間のリンク付け関係を一覧することができる。

【0004】また、特開平5-128157号公報に記載の「文書検索装置」によれば、リンクを辿って到達可能な特定の範囲から、指定したキーワードにマッチする内容をもつ文書を選択的に検索することができる。

【0005】さらに、(株)エーアイソフトの「Web Whacker」(米国ForeFront Group, Inc.の商標)、株式会社ビー・ユー・ジーの「PerMan Surfer波乗野郎」(株式会社ビー・ユー・ジーの商標)などに代表されるオートパイロットあるいはダウンローダと呼ばれるソフトウェアによれば、大規模なハイパーテキストデータベースであるワールドワイドウェブ(World Wide Web: WWW)上の指定された文書からリンクした文書を指定された数だけ、自動収集することができる。

##### 【0006】

【発明が解決しようとする課題】前述の「ハイパーテキストのブラウジング処理装置」では文書間の意味的な関連性は表現されず、ツリー構造状に表示された文書群のどこからどこまでが、意味的に近接した関係にあるのかの判断は、人間が各文書の内容を見て判断するしかない。ワールドワイドウェブのように大規模なハイパーテキストでは、この判断を人手でおこなうのは現実的ではない。

【0007】また、「文書検索装置」を用いると、リンクを辿って到達可能な範囲の文書をすべて一度収集する必要があり、到達可能な範囲が膨大である場合にも、全ての文書の内容を参照する必要があり、莫大な処理時間を要するという問題を生じる。また、通信路の細いネットワーク上に分散された文書を参照する場合には、通信時間などのオーバーヘッドが大きくなるという問題を生じる。キーワード指定がリンクの作成意図と一致しない場合には、キーワードにマッチしない文書を介して間接的にリンクされているキーワードにマッチする文書間でリンクが失われてしまうという問題がある。

【0008】オートパイロットやダウンローダなどのソフトウェアでは、辿るリンクの数や、物理的に文書が格納されているマシンによって文書の収集範囲を限定しており、文書の内容による意味的な関連性は考慮されておらず、内容的にあまり関連のない文書も収集してしまうという問題や、他のマシンに格納されている関連の深い文書が収集されないという問題がある。

【0009】そこで、ワールドワイドウェブのように大規模なハイパーテキストに関しても、リンクの作成意図と文書の内容に沿って関連性の深い文書に限定して収集する方法が必要である。

【0010】本発明の目的は、文書の参照関係に基づいた文書の収集において、到達可能な全文書を探索することなく意味的に関連性の深い文書を収集できるように、参照関係を辿る範囲を限定した収集対象範囲を決定をする文書グループ化装置および文書グループ化方法と文書のグループ化が可能なプログラムを記録した記録媒体を提供することにある。

##### 【0011】

【課題を解決するための手段】本発明の第1の発明の文書グループ化装置は、文書収集の起点となる文書を格納する参照元文書格納手段と、参照元文書格納手段に格納された文書を順次取り出して該文書中から他の文書への参照関係を記述した箇所を検出する参照文書検出手段と、参照文書検出手段により検出された参照関係により、前記参照関係に対応する文書を文書群から取得する文書取得手段と、文書取得手段により取得された文書を格納しておく参照文書格納手段と、参照文書検出手段により文書中の参照関係を検出し終えた参照元文書を参照元文書格納手段から移して格納しておく文書グループ格納手段と、参照文書格納手段に格納された参照文書と文

書グループ格納手段に格納された文書の関連性を評価し、関連がある場合には参照文書を参照元文書格納手段に新たな参照元文書として追加する関連性評価手段とを含んで構成される。

【0012】また、本発明の第2の発明の文書グループ化装置は、文書収集の起点となる文書を格納する参照元文書格納手段と、前記参照元文書格納手段に格納された文書を順次取り出して該文書中から他の文書への参照関係の説明を記述した箇所を検出する参照文書検出手段と、前記参照文書検出手段により文書中の参照関係の説明を検出し終えた参照元文書を前記参照元文書格納手段から移して格納しておく文書グループ格納手段と、前記参照文書検出手段により検出された参照関係の説明に該当する文書中の記述と文書グループ格納手段に格納された参照元文書の本文の記述とから、前記参照元文書と参照関係のある文書との関連性を評価する関連性評価手段と、前記関連性評価手段により関連があると評価された文書のみを文書群から取得する文書取得手段と、前記文書取得手段により取得された文書を格納しておく参照文書格納手段と、から構成され、前記関連性評価手段は、前記参照文書格納手段が格納した文書を参照元文書として、前記参照元文書格納手段に格納する処理を更に行う。

【0013】第1の発明によれば、文書の参照関係に基づいた文書の収集において、到達可能な全文書を探索することなく意味的に関連性の深い文書を収集できるように、参照関係を辿る範囲を限定した収集対象範囲を決定を行うことが可能である。

【0014】また、第2の発明によれば、文書の参照関係の説明によって、文書作成者の意図と文書の内容に沿って関連性の深い文書に限定した収集が可能である。

【0015】

【発明の実施の形態】次に図1から図6を参照して本発明の実施の形態について説明する。

【0016】図1は本発明の第1の発明である請求項1～請求項8に記載した本発明の実施の形態の一構成例を示すブロック図である。

【0017】かかる発明の実施の形態における文書グループ化装置(001)は、文書収集の起点となる文書を格納する参照元文書格納手段(110)と、参照元文書格納手段(110)に格納された文書を順次取り出して該文書中から他の文書への参照関係を検出する参照文書検出手段(120)と、参照文書検出手段(120)により検出された参照関係のある文書を文書群(901)から取得する文書取得手段(130)と、文書取得手段(130)により取得された文書を格納しておく参照文書格納手段(140)と、参照文書検出手段(120)により文書中の参照関係を検出し終えた参照元文書を参照元文書格納手段(110)から移して格納しておく文書グループ格納手段(150)と、参照文書格納手段

(140)に格納された参照文書と文書グループ格納手段(150)に格納された文書群の関連性を評価し、関連がある場合には参照文書を参照元文書格納手段(110)に新たな参照元文書として追加する関連性評価手段(160)とを含んで構成される。

【0018】また、文書グループ格納手段(150)格納されるのは、文書を特定できる情報のみでもよい。

【0019】図2は請求項1から請求項8に記載した本発明の処理の流れの一実施の形態を示すフロー図である。

【0020】参照文書検出手段(120)は、参照元文書格納手段(110)に格納された文書を順次取り出して該文書中から他の文書への参照関係を検出し(ステップS10)、文書中の参照関係を検出し終えた参照元文書を参照元文書格納手段(110)から文書グループ格納手段(150)へ移して格納し(ステップS20)、文書取得手段(130)は、参照文書検出手段(120)により検出された参照関係のある文書を文書群(901)から取得(ステップS30)して参照文書格納手段(140)に格納し、関連性評価手段(160)は、参照文書格納手段(140)に格納された参照文書と文書グループ格納手段(150)に格納された文書群の関連性を評価し(ステップS40)、関連がある場合には参照文書を参照元文書格納手段(110)に新たな参照元文書として追加(ステップS40)し、参照元文書格納手段(110)にまだ文書が格納されているかチェックし(ステップS60)、格納されている場合にはステップS10から繰り返す。参照元文書格納手段(110)格納されている文書が無くなれば、文書グループ格納手段(150)に格納されている文書を一つのグループに属するものと決定する(ステップS70)。

【0021】

【実施例】以下、図面を参照して本発明の文書グループ化装置のさらに詳しい実施例について説明する。

【0022】前述した様に、図1は、本発明の文書グループ化装置の一実施の形態の構成例を示すブロック図である。

【0023】また、本実施例においては、文書群(901)として、HTML形式で記述されているハイパーテキストであり、ワールドワイドウェブのページとしてインターネットに接続された計算機上に分散して存在しているものとして説明する。

【0024】各文書は、通信プロトコルとホスト名およびパス名を含むURL(Uniform Resource Locator)と呼ばれる記述法により特定できる。文書取得手段(130)は、例えばURLに指定されたプロトコルによりインターネットに接続されている指定されたホスト計算機から指定されたパス名に該当する文書を取得する。

【0025】例えば、参照元文書格納手段(110)に

格納された文書収集の起点となる文書が図3に示す文書(501)であるとする。HTML形式の文書では参照関係情報は、"＜"と"＞"に囲まれたタグと呼ばれる部分のうち、"＜a"で始まり次の"＞"までの間にある"href="に続いてURLを記述し、他の文書への参照を示す箇所である。

【0026】参照文書検出手段(120)は、参照元文書中から他の文書への参照を示すURL(参照関係情報)を検出し(ステップS10)、文書取得手段(130)により、そのURLに該当する文書を文書群(901)から取得して、参照文書格納手段(140)に格納する(ステップS30)。図3の文書(501)からは参照文書として、http://www.fisherman.com/maru.htmlおよびhttp://www.shops.com/fishing.htmlの2つのURLが検出される。例えば、これらURLに該当する文書がそれぞれ、図4の文書(502)、図5の文書(503)に示す文書であるとする。文書中のURLを検出し終わると文書(501)は、文書グループ格納手段(150)へ移される(ステップS20)。

【0027】関連性評価手段(160)は例えば、文書グループ格納手段(150)に格納された文書(501)と参照文書格納手段(140)に格納された文書(502)と文書(503)の各文書からタグと不要語を除いてキーワードを抽出し、文書(501)に含まれるキーワードが文書(502)と文書(503)のそれぞれに含まれる度合いを計算して、文書(502)と文書(503)のそれぞれが文書(501)に対する関連性を評価する。

【0028】本例においては、文書(501)の本文中のキーワードが「FISHING、釣り、フライフィッシング」であり、文書(502)のキーワードが「釣り、フライフィッシング、溪流釣り」であり、文書(503)のキーワードが「釣り、ルアー、ショップ」であるとし、関連性を参照元の文書中のキーワード全体に対する参照文書中に含まれる参照元のキーワードの数の比とし、例えば、60%を関連性の有無を判定する基準とすれば、文書(502)の文書(501)に対する関連性は約67%、文書(503)の文書(501)に対する関連性は約33%となり、文書(502)は関連性有り、文書(503)は関連性無しと判定する(ステップS40)。

【0029】関連性無しと判定された文書(503)は、この時点で破棄され、文書(503)からさらに参照される文書があったとしても、それらについては取得しない。

【0030】関連性有りと判定された文書(502)は、参照元文書格納手段(110)に新たな参照元文書として追加し(ステップS60)、以下、文書(50

1)のときと同様に文書(502)を参照元文書として上記の過程を適用し、文書(502)からURLを検出し(ステップS10)、さらに参照される文書を取得する(ステップS30)。

【0031】文書(502)からURLの検出を終え、文書グループ格納手段(150)に格納される(ステップS20)と、文書グループ格納手段(150)には文書(501)と文書(502)の2つの文書が格納されている。関連性評価手段(160)は例えば、これら文書のキーワードの和集合を参照元のキーワード群として、文書(502)から検出されたURLが示す文書の関連性を評価する(ステップS40)。例えばここで、検出されたURLが示す文書がいずれも関連性無しと判定され、参照元文書格納手段に参照元文書がなければ(ステップS60)、処理は終了し(ステップS70)、この時点で文書グループ格納手段(150)に格納されている文書(501)と文書(502)が、ひとつのグループをなす。

【0032】文書グループ格納手段(150)の容量を節約したい場合は、文書を特定するURLとキーワード群のみを文書グループ格納手段(150)に格納してもよい。

【0033】また、関連性評価手段(160)については、このキーワードマッチングによる実施例はあくまで一例であって、本発明は、この実施例だけに限定されるものではない。例えば、シソーラスなどを用いてキーワード間の距離を計算し、参照される文書のキーワード群の間の距離の総和や平均を用い、距離の大きさを判定基準に用いることもできる。

【0034】次に、本発明の第2の発明である請求項9～12に記載した発明を図面を参照して説明する。

【0035】図6は、第2の発明の実施の形態の一構成例を示すブロック図である。本実施例においては、対象文書としてHTML形式のハイパーテキスト文書を扱う場合の実施例について説明する。また、先に説明した第1の発明と機能が重なる箇所については、説明を省略する。

【0036】第2の発明の関連性評価手段(160)は、参照文書検出手段(120)により検出されるURLを説明している参照元文書中の文字列と、文書グループ格納手段(150)中の文書からその文字列を除いた部分との関連性を判定することで、参照される文書の関連性を推定し、関連性があると推定された文書についてのみ文書取得手段(130)により、そのURLに該当する文書を文書群(901)から取得して、参照文書格納手段(140)に格納する。

【0037】例えば、図3の文書(501)では、URLが記述されているタグ"＜a…＞"と対応するタグ"＜/a＞"との間の文字列をURLに対する説明となる文字列とみなし、URL「http://www.f



fisherman.com/maru.html」に対して「丸山さんの釣り情報(フライフィッシングの話題もあり)」という文字列が、この参照文書(URL)を説明する文書になり、さらにURL「http://www.shops.com/fishing.html」に対して「その他の釣り情報」という文字列がこの参照文書(URL)を説明する文書となる。それぞれの文字列のキーワードは「丸山、釣り、フライフィッシング」、「その他、釣り」となる。

【0038】ここで、前述した様な、参照元文書と参照文書間と同様な関連性判定を行えば、URL「http://www.fisherman.com/maru.html」の説明のキーワード「丸山、釣り、フライフィッシング」が、文書(501)の本文中に含まれている率(関連性)は約67%となり、URL「http://www.shops.com/fishing.html」の説明のキーワード「その他、釣り」が文書(501)の本文中に含まれている率(関連性)は50%となる。ここで、60%を関連性の有り無しを判定する基準とすれば、URL「http://www.fisherman.com/maru.html」は関連性有り、URL「http://www.shops.com/fishing.html」は関連性無しと判定する。

【0039】関連性無しと判定されたURL「http://www.shops.com/fishing.html」の文書(503)の取得はおこなわず、URL「http://www.fisherman.com/maru.html」の文書(502)のみを文書取得手段(130)により取得し、参照文書格納手段(140)に格納する。その後は前述した実施例の説明と同様に文書間の関連性判定をおこなって処理を継続する。あるいはURLの説明による関連性の推定を信頼して、文書間の関連性判定を省略することもできる。

【図3】

```

<html>
<head>
<title>MY HOBBY</title>
</head>
<body>
FISHING
</body>
</html>
最近の私の釣りはフライフィッシングばかりです。
<a href="http://www.fisherman.com/maru.html">
丸山さんの釣り情報(フライフィッシングの話題もあり)</a>
<a href="http://www.shops.com/fishing.html">
その他の釣り情報</a>
</body>
</html>

```

【0040】また、本発明においては、以上の述べたような構成をコンピュータプログラムによって作成し、フロッピーディスクやCD-ROMに代表される記録媒体によって記録してもよい。

【0041】

【発明の効果】本発明によれば、文書の参照関係に基づいた文書の収集において、到達可能な全文書を探索することなく意味的に関連性の深い文書を収集できるように、参照関係を辿る範囲を限定した収集対象範囲を決定をする文書グループ化装置を提供でき、ワールドワイドウェブのように大規模なハイパーテキストに関しても、リンクの作成意図と文書の内容に沿って関連性の深い文書に限定した収集が可能になる。

【図面の簡単な説明】

【図1】本発明の文書グループ化装置の実施の形態の一構成例を示すブロック図

【図2】本発明の文書グループ化装置の処理の流れの一実施の形態を示すフロー図

【図3】HTML形式の文書の一例を示す図

【図4】HTML形式の文書の一例を示す図

【図5】HTML形式の文書の一例を示す図

【図6】本発明の文書グループ化装置の実施の形態の他の構成例を示すブロック図。

【符号の説明】

001 文書グループ化装置

110 参照元文書格納手段

120 参照文書検出手段

130 文書取得手段

140 参照文書格納手段

150 文書グループ格納手段

160 関連性評価手段

501、502、503 HTML形式の文書の例

901 文書群

【図4】

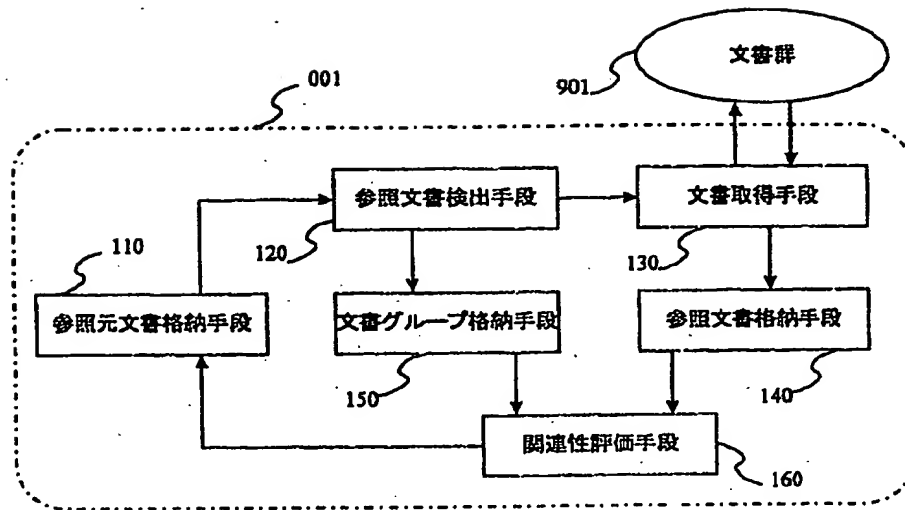
```

<html>
<title>釣り丸山のページ</title>
</head>
<body>
<h1>
<h2>最近の釣り
<h3>おすすめは三角湖の釣り場のもと。
<h4>フライフィッシング
<h5>最近では黒川の下流が好過。
</h5>
<a href="http://www.shops.com">近所の釣り具屋</a>
</body>
</html>

```



【図1】



【図5】

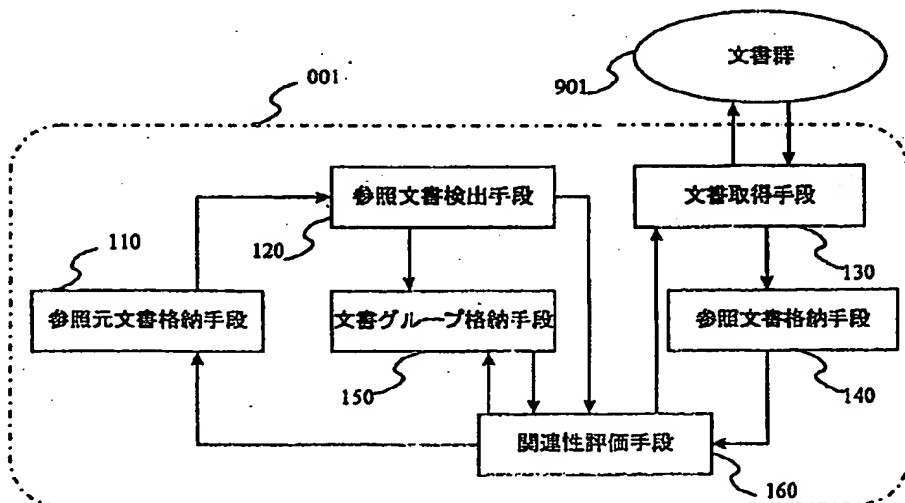
503

```

<html>
<head>
<title>のり関連ショップ</title>
</head>
<body>
ルアーの組立<br>
<a href=daicho.html>フィッシング「ショップ大船」</a><br>
<a href=izumi.html>のり風の「いれぐい」</a><br>
</body>
</html>

```

【図6】



【図2】

